Cyclistic bike-share analysis case study

Ugochukwu Orji

11/9/2021

Data Analysis method:

Ask -> Prepare -> Process -> Analyze -> Share -> Act

1. Ask

Background

• Company: Cyclistic, a fictional bike-share company based in Chicago launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geo-tracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime. Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

• My role: Junior data analyst working in the marketing analytics team

• Premise: Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. The director of marketing (who is also my manager) believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs. The director of marketing wants to Design marketing strategies aimed at converting casual riders into annual members but to do this, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. So there's need to analyze the Cyclistic historical bike trip data to identify trends

The Business Task:

My task is to analyze the Cyclistic historical bike trip data to identify trends that will help the marketing team to convert casual riders into members.

Key Stakeholders

• Primary stakeholder - the director of marketing (my manager)

• Secondary stakeholders - the marketing analytics team and the executive team

3. Process

Reading the packages into R

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(skimr)
library(here)
```

```
## here() starts at C:/Users/hp/Documents/Google capstone
```

```
library(dplyr)
library(modeest)
```

```
## Registered S3 method overwritten by 'rmutil':
##   method         from
##   print.response httr
```

```
library(DescTools)
```

Setting file directory

```
setwd("~/Google capstone/Capstone markdown") # Set working directory
getwd() # displays your working directory
```

```
## [1] "C:/Users/hp/Documents/Google capstone/Capstone markdown"
```

STEP 1: COLLECT DATA

Data Source

Previous 12 months (from Oct, 2020 to Sept, 2021) of Cyclistic trip data downloaded here.

Important Notes

- Cyclistic is a fictional company thus the datasets name do not match
- The dataset is licensed under this regulation and has been publicly made available by Motivate International Inc.

Uploading Divvy datasets (csv files) here

```
Oct_2020_tripdata <- read_csv("Oct 2020.csv")
```

```
## Rows: 388653 Columns: 13
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): ride_id, rideable_type, started_at, ended_at, start_station_name, e...
## dbl (6): start_station_id, end_station_id, start_lat, start_lng, end_lat, en...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Nov_2020_tripdata <- read_csv("Nov 2020.csv")
```

```
## Rows: 259716 Columns: 13
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): ride_id, rideable_type, started_at, ended_at, start_station_name, e...
## dbl (6): start_station_id, end_station_id, start_lat, start_lng, end_lat, en...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Dec_2020_tripdata <- read_csv("Dec 2020.csv")
```

```
## Rows: 131573 Columns: 13
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): ride_id, rideable_type, started_at, ended_at, start_station_name, s...
## dbl (4): start_lat, start_lng, end_lat, end_lng
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Jan_2021_tripdata <- read_csv("Jan 2021.csv")
```

```
## Rows: 96834 Columns: 13
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): ride_id, rideable_type, started_at, ended_at, start_station_name, s...
```

```
## dbl (4): start_lat, start_lng, end_lat, end_lng

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Feb_2021_tripdata <- read_csv("Feb 2021.csv")
```

```
## Rows: 49622 Columns: 13

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): ride_id, rideable_type, started_at, ended_at, start_station_name, s...
## dbl (4): start_lat, start_lng, end_lat, end_lng

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Mar_2021_tripdata <- read_csv("Mar 2021.csv")
```

```
## Rows: 228496 Columns: 13

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): ride_id, rideable_type, started_at, ended_at, start_station_name, s...
## dbl (4): start_lat, start_lng, end_lat, end_lng

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Apr_2021_tripdata <- read_csv("Apr 2021.csv")
```

```
## Rows: 337230 Columns: 13

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): ride_id, rideable_type, started_at, ended_at, start_station_name, s...
## dbl (4): start_lat, start_lng, end_lat, end_lng

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
May_2021_tripdata <- read_csv("May 2021.csv")
```

```
## Rows: 531633 Columns: 13

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): ride_id, rideable_type, started_at, ended_at, start_station_name, s...
## dbl (4): start_lat, start_lng, end_lat, end_lng

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Jun_2021_tripdata <- read_csv("Jun 2021.csv")
```

```
## Rows: 729595 Columns: 13
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): ride_id, rideable_type, started_at, ended_at, start_station_name, s...
## dbl (4): start_lat, start_lng, end_lat, end_lng

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Jul_2021_tripdata <- read_csv("Jul 2021.csv")
```

```
## Rows: 822410 Columns: 13

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): ride_id, rideable_type, started_at, ended_at, start_station_name, s...
## dbl (4): start_lat, start_lng, end_lat, end_lng

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Aug_2021_tripdata <- read_csv("Aug 2021.csv")
```

```
## Rows: 804352 Columns: 13

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): ride_id, rideable_type, started_at, ended_at, start_station_name, s...
## dbl (4): start_lat, start_lng, end_lat, end_lng

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Sept_2021_tripdata <- read_csv("Sept 2021.csv")
```

```
## Rows: 756147 Columns: 13

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): ride_id, rideable_type, started_at, ended_at, start_station_name, s...
## dbl (4): start_lat, start_lng, end_lat, end_lng

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

2. Prepare

The dataset follows the ROCCC Analysis as described below:

- Reliable - yes, not biased
- Original - yes, can locate the original public data
- Comprehensive - yes, not missing important information
- Current - yes, updated monthly
- Cited - yes

STEP 2: WRANGLE DATA AND COMBINE INTO A SINGLE FILE

Display each column name and check for consistency

Inspect the dataframes and look for inconsistencies

```
str(Oct_2020_tripdata)
```

```
## spec_tbl_df [388,653 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:388653] "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396B54A15AC0DF" "44A4A
##  $ rideable_type     : chr [1:388653] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : chr [1:388653] "10/31/2020 19:39" "10/31/2020 23:50" "10/31/2020 23:00" "10/3
##  $ ended_at          : chr [1:388653] "10/31/2020 19:57" "11/1/2020 0:04" "10/31/2020 23:08" "10/31/
##  $ start_station_name: chr [1:388653] "Lakeview Ave & Fullerton Pkwy" "Southport Ave & Waveland Ave"
##  $ start_station_id  : num [1:388653] 313 227 102 165 190 359 313 125 NA 174 ...
##  $ end_station_name  : chr [1:388653] "Rush St & Hubbard St" "Kedzie Ave & Milwaukee Ave" "University
##  $ end_station_id    : num [1:388653] 125 260 423 256 185 53 125 313 199 635 ...
##  $ start_lat         : num [1:388653] 41.9 41.9 41.8 42 41.9 ...
##  $ start_lng         : num [1:388653] -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num [1:388653] 41.9 41.9 41.8 42 41.9 ...
##  $ end_lng           : num [1:388653] -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:388653] "casual" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_character(),
##   ..   ended_at = col_character(),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_double(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_double(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(Nov_2020_tripdata)
```

```
## spec_tbl_df [259,716 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:259716] "BD0A6FF6FFF9B921" "96A7A7A4BDE4F82D" "C61526D06582BDC5" "E533
##  $ rideable_type     : chr [1:259716] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : chr [1:259716] "11/1/2020 13:36" "11/1/2020 10:03" "11/1/2020 0:34" "11/1/202
##  $ ended_at          : chr [1:259716] "11/1/2020 13:45" "11/1/2020 10:14" "11/1/2020 1:03" "11/1/202
##  $ start_station_name: chr [1:259716] "Dearborn St & Erie St" "Franklin St & Illinois St" "Lake Shor
##  $ start_station_id  : num [1:259716] 110 672 76 659 2 72 76 NA 58 394 ...
##  $ end_station_name  : chr [1:259716] "St. Clair St & Erie St" "Noble St & Milwaukee Ave" "Federal S
##  $ end_station_id    : num [1:259716] 211 29 41 185 2 76 72 NA 288 273 ...
##  $ start_lat         : num [1:259716] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:259716] -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:259716] 41.9 41.9 41.9 41.9 41.9 ...
```

```
##  $ end_lng           : num [1:259716] -87.6 -87.7 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr [1:259716] "casual" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_character(),
##   ..   ended_at = col_character(),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_double(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_double(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(Dec_2020_tripdata)
```

```
## spec_tbl_df [131,573 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:131573] "70B6A9A437D4C30D" "158A465D4E74C54A" "5262016E0F1F2F9A" "BE11
##  $ rideable_type     : chr [1:131573] "classic_bike" "electric_bike" "electric_bike" "electric_bike"
##  $ started_at        : chr [1:131573] "12/27/2020 12:44" "12/18/2020 17:37" "12/15/2020 15:04" "12/15
##  $ ended_at          : chr [1:131573] "12/27/2020 12:55" "12/18/2020 17:44" "12/15/2020 15:11" "12/15
##  $ start_station_name: chr [1:131573] "Aberdeen St & Jackson Blvd" NA NA NA ...
##  $ start_station_id  : chr [1:131573] "13157" NA NA NA ...
##  $ end_station_name  : chr [1:131573] "Desplaines St & Kinzie St" NA NA NA ...
##  $ end_station_id    : chr [1:131573] "TA1306000003" NA NA NA ...
##  $ start_lat         : num [1:131573] 41.9 41.9 41.9 41.9 41.8 ...
##  $ start_lng         : num [1:131573] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:131573] 41.9 41.9 41.9 41.9 41.8 ...
##  $ end_lng           : num [1:131573] -87.6 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr [1:131573] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_character(),
##   ..   ended_at = col_character(),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(Jan_2021_tripdata)
```

```
## spec_tbl_df [96,834 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:96834] "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F27" "4FA453
##  $ rideable_type     : chr [1:96834] "electric_bike" "electric_bike" "electric_bike" "electric_bike"
##  $ started_at        : chr [1:96834] "1/23/2021 16:14" "1/27/2021 18:43" "1/21/2021 22:35" "1/7/2021
##  $ ended_at          : chr [1:96834] "1/23/2021 16:24" "1/27/2021 18:47" "1/21/2021 22:37" "1/7/2021
##  $ start_station_name: chr [1:96834] "California Ave & Cortez St" "California Ave & Cortez St" "Cali
##  $ start_station_id  : chr [1:96834] "17660" "17660" "17660" "17660" ...
##  $ end_station_name  : chr [1:96834] NA NA NA NA ...
##  $ end_station_id    : chr [1:96834] NA NA NA NA ...
##  $ start_lat         : num [1:96834] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:96834] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:96834] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:96834] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:96834] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_character(),
##   ..   ended_at = col_character(),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(Feb_2021_tripdata)
```

```
## spec_tbl_df [49,622 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:49622] "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91" "B32D3
##  $ rideable_type     : chr [1:49622] "classic_bike" "classic_bike" "electric_bike" "classic_bike" ..
##  $ started_at        : chr [1:49622] "2/12/2021 16:14" "2/14/2021 17:52" "2/9/2021 19:10" "2/2/2021
##  $ ended_at          : chr [1:49622] "2/12/2021 16:21" "2/14/2021 18:12" "2/9/2021 19:19" "2/2/2021
##  $ start_station_name: chr [1:49622] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Clark St
##  $ start_station_id  : chr [1:49622] "525" "525" "KA1503000012" "637" ...
##  $ end_station_name  : chr [1:49622] "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St" "State S
##  $ end_station_id    : chr [1:49622] "660" "16806" "TA1305000029" "TA1305000034" ...
##  $ start_lat         : num [1:49622] 42 42 41.9 41.9 41.8 ...
##  $ start_lng         : num [1:49622] -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:49622] 42 42 41.9 41.9 41.8 ...
##  $ end_lng           : num [1:49622] -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr [1:49622] "member" "casual" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
```

```
##   ..   started_at = col_character(),
##   ..   ended_at = col_character(),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(Mar_2021_tripdata)
```

```
## spec_tbl_df [228,496 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:228496] "CFA86D4455AA1030" "30D9DC61227D1AF3" "846D87A15682A284" "994D
##  $ rideable_type     : chr [1:228496] "classic_bike" "classic_bike" "classic_bike" "classic_bike" ..
##  $ started_at        : chr [1:228496] "3/16/2021 8:32" "3/28/2021 1:26" "3/11/2021 21:17" "3/11/2021
##  $ ended_at          : chr [1:228496] "3/16/2021 8:36" "3/28/2021 1:36" "3/11/2021 21:33" "3/11/2021
##  $ start_station_name: chr [1:228496] "Humboldt Blvd & Armitage Ave" "Humboldt Blvd & Armitage Ave"
##  $ start_station_id  : chr [1:228496] "15651" "15651" "15443" "TA1308000021" ...
##  $ end_station_name  : chr [1:228496] "Stave St & Armitage Ave" "Central Park Ave & Bloomingdale Ave
##  $ end_station_id    : chr [1:228496] "13266" "18017" "TA1308000043" "13323" ...
##  $ start_lat         : num [1:228496] 41.9 41.9 41.8 42 42 ...
##  $ start_lng         : num [1:228496] -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num [1:228496] 41.9 41.9 41.8 42 42.1 ...
##  $ end_lng           : num [1:228496] -87.7 -87.7 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr [1:228496] "casual" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_character(),
##   ..   ended_at = col_character(),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(Apr_2021_tripdata)
```

```
## spec_tbl_df [337,230 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:337230] "6C992BD37A98A63F" "1E0145613A209000" "E498E15508A80BAD" "1887
##  $ rideable_type     : chr [1:337230] "classic_bike" "docked_bike" "docked_bike" "classic_bike" ...
##  $ started_at        : chr [1:337230] "4/12/2021 18:25" "4/27/2021 17:27" "4/3/2021 12:42" "4/17/202
##  $ ended_at          : chr [1:337230] "4/12/2021 18:56" "4/27/2021 18:31" "4/7/2021 11:40" "4/17/202
```

```
##  $ start_station_name: chr [1:337230] "State St & Pearson St" "Dorchester Ave & 49th St" "Loomis Blvd
##  $ start_station_id  : chr [1:337230] "TA1307000061" "KA1503000069" "20121" "TA1305000034" ...
##  $ end_station_name  : chr [1:337230] "Southport Ave & Waveland Ave" "Dorchester Ave & 49th St" "Loom
##  $ end_station_id    : chr [1:337230] "13235" "KA1503000069" "20121" "13235" ...
##  $ start_lat         : num [1:337230] 41.9 41.8 41.7 41.9 41.7 ...
##  $ start_lng         : num [1:337230] -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:337230] 41.9 41.8 41.7 41.9 41.7 ...
##  $ end_lng           : num [1:337230] -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:337230] "member" "casual" "casual" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_character(),
##   ..   ended_at = col_character(),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(May_2021_tripdata)
```

```
## spec_tbl_df [531,633 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:531633] "C809ED75D6160B2A" "DD59FDCE0ACACAF3" "0AB83CB88C43EFC2" "7881A
##  $ rideable_type     : chr [1:531633] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : chr [1:531633] "5/30/2021 11:58" "5/30/2021 11:29" "5/30/2021 14:24" "5/30/202
##  $ ended_at          : chr [1:531633] "5/30/2021 12:10" "5/30/2021 12:14" "5/30/2021 14:25" "5/30/202
##  $ start_station_name: chr [1:531633] NA NA NA NA ...
##  $ start_station_id  : chr [1:531633] NA NA NA NA ...
##  $ end_station_name  : chr [1:531633] NA NA NA NA ...
##  $ end_station_id    : chr [1:531633] NA NA NA NA ...
##  $ start_lat         : num [1:531633] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:531633] -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:531633] 41.9 41.8 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:531633] -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:531633] "casual" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_character(),
##   ..   ended_at = col_character(),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
```

```
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(Jun_2021_tripdata)
```

```
## spec_tbl_df [729,595 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:729595] "99FEC93BA843FB20" "06048DCFC8520CAF" "9598066F68045DF2" "B03C
##  $ rideable_type     : chr [1:729595] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : chr [1:729595] "6/13/2021 14:31" "6/4/2021 11:18" "6/4/2021 9:49" "6/3/2021 19
##  $ ended_at          : chr [1:729595] "6/13/2021 14:34" "6/4/2021 11:24" "6/4/2021 9:55" "6/3/2021 20
##  $ start_station_name: chr [1:729595] NA NA NA NA ...
##  $ start_station_id  : chr [1:729595] NA NA NA NA ...
##  $ end_station_name  : chr [1:729595] NA NA NA NA ...
##  $ end_station_id    : chr [1:729595] NA NA NA NA ...
##  $ start_lat         : num [1:729595] 41.8 41.8 41.8 41.8 41.8 ...
##  $ start_lng         : num [1:729595] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num [1:729595] 41.8 41.8 41.8 41.8 41.8 ...
##  $ end_lng           : num [1:729595] -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr [1:729595] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_character(),
##   ..   ended_at = col_character(),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(Jul_2021_tripdata)
```

```
## spec_tbl_df [822,410 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:822410] "0A1B623926EF4E16" "B2D5583A5A5E76EE" "6F264597DDBF427A" "379B5
##  $ rideable_type     : chr [1:822410] "docked_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : chr [1:822410] "7/2/2021 14:44" "7/7/2021 16:57" "7/25/2021 11:30" "7/8/2021 1
##  $ ended_at          : chr [1:822410] "7/2/2021 15:19" "7/7/2021 17:16" "7/25/2021 11:48" "7/8/2021 1
##  $ start_station_name: chr [1:822410] "Michigan Ave & Washington St" "California Ave & Cortez St" "Wa
##  $ start_station_id  : chr [1:822410] "13001" "17660" "SL-012" "17660" ...
##  $ end_station_name  : chr [1:822410] "Halsted St & North Branch St" "Wood St & Hubbard St" "Rush St
##  $ end_station_id    : chr [1:822410] "KA1504000117" "13432" "KA1503000044" "13196" ...
##  $ start_lat         : num [1:822410] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:822410] -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num [1:822410] 41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num [1:822410] -87.6 -87.7 -87.6 -87.7 -87.7 ...
```

```
##  $ member_casual     : chr [1:822410] "casual" "casual" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_character(),
##   ..   ended_at = col_character(),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(Aug_2021_tripdata)
```

```
## spec_tbl_df [804,352 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:804352] "99103BB87CC6C1BB" "EAFCCCFB0A3FC5A1" "9EF4F46C57AD234D" "5834
##  $ rideable_type     : chr [1:804352] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : chr [1:804352] "8/10/2021 17:15" "8/10/2021 17:23" "8/21/2021 2:34" "8/21/202
##  $ ended_at          : chr [1:804352] "8/10/2021 17:22" "8/10/2021 17:39" "8/21/2021 2:50" "8/21/202
##  $ start_station_name: chr [1:804352] NA NA NA NA ...
##  $ start_station_id  : chr [1:804352] NA NA NA NA ...
##  $ end_station_name  : chr [1:804352] NA NA NA NA ...
##  $ end_station_id    : chr [1:804352] NA NA NA NA ...
##  $ start_lat         : num [1:804352] 41.8 41.8 42 42 41.8 ...
##  $ start_lng         : num [1:804352] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:804352] 41.8 41.8 42 42 41.8 ...
##  $ end_lng           : num [1:804352] -87.7 -87.6 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr [1:804352] "member" "member" "member" "member" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_character(),
##   ..   ended_at = col_character(),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(Sept_2021_tripdata)
```

```
## spec_tbl_df [756,147 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:756147] "9DC7B962304CBFD8" "F930E2C6872D6B32" "6EF72137900BB910" "78D1
##  $ rideable_type     : chr [1:756147] "electric_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : chr [1:756147] "9/28/2021 16:07" "9/28/2021 14:24" "9/28/2021 0:20" "9/28/202
##  $ ended_at          : chr [1:756147] "9/28/2021 16:09" "9/28/2021 14:40" "9/28/2021 0:23" "9/28/202
##  $ start_station_name: chr [1:756147] NA NA NA NA ...
##  $ start_station_id  : chr [1:756147] NA NA NA NA ...
##  $ end_station_name  : chr [1:756147] NA NA NA NA ...
##  $ end_station_id    : chr [1:756147] NA NA NA NA ...
##  $ start_lat         : num [1:756147] 41.9 41.9 41.8 41.8 41.9 ...
##  $ start_lng         : num [1:756147] -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:756147] 41.9 42 41.8 41.8 41.9 ...
##  $ end_lng           : num [1:756147] -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr [1:756147] "casual" "casual" "casual" "casual" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_character(),
##   ..   ended_at = col_character(),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Converting relevant so that they can stack correctly

```
Oct_2020_tripdata <-  mutate(Oct_2020_tripdata, start_station_id = as.character(start_station_id)
                             ,end_station_id = as.character(end_station_id)
                             ,started_at= as.POSIXct(started_at, format= "%m/%d/%Y %H:%M")
                             ,ended_at= as.POSIXct(ended_at, format= "%m/%d/%Y %H:%M"))
Nov_2020_tripdata <-  mutate(Nov_2020_tripdata, start_station_id = as.character(start_station_id)
                             ,end_station_id = as.character(end_station_id)
                             ,started_at= as.POSIXct(started_at, format= "%m/%d/%Y %H:%M")
                             ,ended_at= as.POSIXct(ended_at, format= "%m/%d/%Y %H:%M"))
Dec_2020_tripdata <-  mutate(Dec_2020_tripdata, start_station_id = as.character(start_station_id)
                             ,end_station_id = as.character(end_station_id)
                             ,started_at= as.POSIXct(started_at, format= "%m/%d/%Y %H:%M")
                             ,ended_at= as.POSIXct(ended_at, format= "%m/%d/%Y %H:%M"))
Jan_2021_tripdata <-  mutate(Jan_2021_tripdata, start_station_id = as.character(start_station_id)
                             ,end_station_id = as.character(end_station_id)
                             ,started_at= as.POSIXct(started_at, format= "%m/%d/%Y %H:%M")
                             ,ended_at= as.POSIXct(ended_at, format= "%m/%d/%Y %H:%M"))
Feb_2021_tripdata <-  mutate(Feb_2021_tripdata, start_station_id = as.character(start_station_id)
```

```
                            ,end_station_id = as.character(end_station_id)
                            ,started_at= as.POSIXct(started_at, format= "%m/%d/%Y %H:%M")
                            ,ended_at= as.POSIXct(ended_at, format= "%m/%d/%Y %H:%M"))
Mar_2021_tripdata <-  mutate(Mar_2021_tripdata, start_station_id = as.character(start_station_id)
                            ,end_station_id = as.character(end_station_id)
                            ,started_at= as.POSIXct(started_at, format= "%m/%d/%Y %H:%M")
                            ,ended_at= as.POSIXct(ended_at, format= "%m/%d/%Y %H:%M"))
Apr_2021_tripdata <-  mutate(Apr_2021_tripdata, start_station_id = as.character(start_station_id)
                            ,end_station_id = as.character(end_station_id)
                            ,started_at= as.POSIXct(started_at, format= "%m/%d/%Y %H:%M")
                            ,ended_at= as.POSIXct(ended_at, format= "%m/%d/%Y %H:%M"))
May_2021_tripdata <-  mutate(May_2021_tripdata, start_station_id = as.character(start_station_id)
                            ,end_station_id = as.character(end_station_id)
                            ,started_at= as.POSIXct(started_at, format= "%m/%d/%Y %H:%M")
                            ,ended_at= as.POSIXct(ended_at, format= "%m/%d/%Y %H:%M"))
Jun_2021_tripdata <-  mutate(Jun_2021_tripdata, start_station_id = as.character(start_station_id)
                            ,end_station_id = as.character(end_station_id)
                            ,started_at= as.POSIXct(started_at, format= "%m/%d/%Y %H:%M")
                            ,ended_at= as.POSIXct(ended_at, format= "%m/%d/%Y %H:%M"))
Jul_2021_tripdata <-  mutate(Jul_2021_tripdata, start_station_id = as.character(start_station_id)
                            ,end_station_id = as.character(end_station_id)
                            ,started_at= as.POSIXct(started_at, format= "%m/%d/%Y %H:%M")
                            ,ended_at= as.POSIXct(ended_at, format= "%m/%d/%Y %H:%M"))
Aug_2021_tripdata <-  mutate(Aug_2021_tripdata, start_station_id = as.character(start_station_id)
                            ,end_station_id = as.character(end_station_id)
                            ,started_at= as.POSIXct(started_at, format= "%m/%d/%Y %H:%M")
                            ,ended_at= as.POSIXct(ended_at, format= "%m/%d/%Y %H:%M"))
Sept_2021_tripdata <-  mutate(Sept_2021_tripdata, start_station_id = as.character(start_station_id)
                            ,end_station_id = as.character(end_station_id)
                            ,started_at= as.POSIXct(started_at, format= "%m/%d/%Y %H:%M")
                            ,ended_at= as.POSIXct(ended_at, format= "%m/%d/%Y %H:%M"))
```

Binding data frames into one big data frame

```
all_trips <- bind_rows(Oct_2020_tripdata, Nov_2020_tripdata, Dec_2020_tripdata,
                       Jan_2021_tripdata, Feb_2021_tripdata, Mar_2021_tripdata,
                       Apr_2021_tripdata, May_2021_tripdata, Jun_2021_tripdata,
                       Jul_2021_tripdata, Aug_2021_tripdata, Sept_2021_tripdata)
```

Remove lat, long

```
all_trips <- all_trips %>%
  select(-c(start_lat, start_lng, end_lat, end_lng))
```

STEP 3: CLEAN UP AND ADD DATA TO PREPARE FOR ANALYSIS

Inspect the new table that has been created

```
colnames(all_trips)  #List of column names
```

```
## [1] "ride_id"            "rideable_type"      "started_at"
## [4] "ended_at"           "start_station_name" "start_station_id"
## [7] "end_station_name"   "end_station_id"     "member_casual"
```

```
nrow(all_trips)  #How many rows are in data frame?
```

```
## [1] 5136261
```

```
dim(all_trips)  #Dimensions of the data frame?
```

```
## [1] 5136261       9
```

```
head(all_trips)  #See the first 6 rows of data frame.  Also tail(all_trips)
```

```
## # A tibble: 6 x 9
##   ride_id          rideable_type started_at          ended_at            start_station_n~
##   <chr>            <chr>         <dttm>              <dttm>              <chr>
## 1 ACB6B40CF5B9044C electric_bike 2020-10-31 19:39:00 2020-10-31 19:57:00 Lakeview Ave & ~
## 2 DF450C72FD109C01 electric_bike 2020-10-31 23:50:00 2020-11-01 00:04:00 Southport Ave &~
## 3 B6396B54A15AC0DF electric_bike 2020-10-31 23:00:00 2020-10-31 23:08:00 Stony Island Av~
## 4 44A4AEE261B9E854 electric_bike 2020-10-31 22:16:00 2020-10-31 22:19:00 Clark St & Grac~
## 5 10B7DD76A6A2EB95 electric_bike 2020-10-31 19:38:00 2020-10-31 19:54:00 Southport Ave &~
## 6 DA6C3759660133DA electric_bike 2020-10-29 17:38:00 2020-10-29 17:45:00 Larrabee St & D~
## # ... with 4 more variables: start_station_id <chr>, end_station_name <chr>,
## #   end_station_id <chr>, member_casual <chr>
```

```
str(all_trips)  #See list of columns and data types (numeric, character, etc)
```

```
## tibble [5,136,261 x 9] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5136261] "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396B54A15AC0DF" "44A4
##  $ rideable_type     : chr [1:5136261] "electric_bike" "electric_bike" "electric_bike" "electric_bik
##  $ started_at        : POSIXct[1:5136261], format: "2020-10-31 19:39:00" "2020-10-31 23:50:00" ...
##  $ ended_at          : POSIXct[1:5136261], format: "2020-10-31 19:57:00" "2020-11-01 00:04:00" ...
##  $ start_station_name: chr [1:5136261] "Lakeview Ave & Fullerton Pkwy" "Southport Ave & Waveland Ave
##  $ start_station_id  : chr [1:5136261] "313" "227" "102" "165" ...
##  $ end_station_name  : chr [1:5136261] "Rush St & Hubbard St" "Kedzie Ave & Milwaukee Ave" "Universi
##  $ end_station_id    : chr [1:5136261] "125" "260" "423" "256" ...
##  $ member_casual     : chr [1:5136261] "casual" "casual" "casual" "casual" ...
```

```
summary(all_trips)  #Statistical summary of data. Mainly for numerics
```

```
##    ride_id          rideable_type        started_at
##  Length:5136261     Length:5136261     Min.   :2020-10-01 00:00:00
##  Class :character   Class :character   1st Qu.:2021-04-11 18:50:00
##  Mode  :character   Mode  :character   Median :2021-06-21 18:01:00
##                                        Mean   :2021-05-25 22:30:27
##                                        3rd Qu.:2021-08-11 21:13:00
##                                        Max.   :2021-09-30 23:59:00
##     ended_at                   start_station_name start_station_id
##  Min.   :2020-10-01 00:05:00   Length:5136261     Length:5136261
##  1st Qu.:2021-04-11 19:15:00   Class :character   Class :character
##  Median :2021-06-21 18:20:00   Mode  :character   Mode  :character
##  Mean   :2021-05-25 22:51:05
##  3rd Qu.:2021-08-11 21:33:00
##  Max.   :2021-10-01 22:55:00
##  end_station_name   end_station_id     member_casual
##  Length:5136261     Length:5136261     Length:5136261
##  Class :character   Class :character   Class :character
```

```
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Display some elements of the new table to see if everything is as expected

```
glimpse(all_trips)
```

```
## Rows: 5,136,261
## Columns: 9
## $ ride_id            <chr> "ACB6B40CF5B9044C", "DF450C72FD109C01", "B6396B54A1~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <dttm> 2020-10-31 19:39:00, 2020-10-31 23:50:00, 2020-10-~
## $ ended_at           <dttm> 2020-10-31 19:57:00, 2020-11-01 00:04:00, 2020-10-~
## $ start_station_name <chr> "Lakeview Ave & Fullerton Pkwy", "Southport Ave & W~
## $ start_station_id   <chr> "313", "227", "102", "165", "190", "359", "313", "1~
## $ end_station_name   <chr> "Rush St & Hubbard St", "Kedzie Ave & Milwaukee Ave~
## $ end_station_id     <chr> "125", "260", "423", "256", "185", "53", "125", "31~
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "~
```

Removing rows with missing values

```
colSums(is.na(all_trips))
```

```
##            ride_id      rideable_type         started_at           ended_at
##                  0                  0                  0                  0
## start_station_name   start_station_id   end_station_name     end_station_id
##             523467             523781             567268             567501
##      member_casual
##                  0
```

```
all_trips_cleaned <- all_trips[complete.cases(all_trips), ]
```

Flitering started_at data that is greater than ended_at

```
all_trips_cleaned <- all_trips_cleaned %>%
  filter(all_trips_cleaned$started_at < all_trips_cleaned$ended_at)
```

New columns to list the date, month, day, and year of each ride

```
all_trips_cleaned$date <- as.Date(all_trips_cleaned$started_at, format= "%m/%d/%Y %H:%M")
all_trips_cleaned$month <- format(as.Date(all_trips_cleaned$date), "%m")
all_trips_cleaned$day <- format(as.Date(all_trips_cleaned$date), "%d")
all_trips_cleaned$year <- format(as.Date(all_trips_cleaned$date), "%Y")
all_trips_cleaned$day_of_week <- format(as.Date(all_trips_cleaned$date), "%A")
```

Display some elements of the new table to see if everything is as expected

```
glimpse(all_trips_cleaned)
```

```
## Rows: 4,317,599
## Columns: 14
```

```
## $ ride_id            <chr> "ACB6B40CF5B9044C", "DF450C72FD109C01", "B6396B54A1~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <dttm> 2020-10-31 19:39:00, 2020-10-31 23:50:00, 2020-10-~
## $ ended_at           <dttm> 2020-10-31 19:57:00, 2020-11-01 00:04:00, 2020-10-~
## $ start_station_name <chr> "Lakeview Ave & Fullerton Pkwy", "Southport Ave & W~
## $ start_station_id   <chr> "313", "227", "102", "165", "190", "359", "313", "1~
## $ end_station_name   <chr> "Rush St & Hubbard St", "Kedzie Ave & Milwaukee Ave~
## $ end_station_id     <chr> "125", "260", "423", "256", "185", "53", "125", "31~
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "~
## $ date               <date> 2020-10-31, 2020-10-31, 2020-10-31, 2020-10-31, 20~
## $ month              <chr> "10", "10", "10", "10", "10", "10", "10", "10", "10~
## $ day                <chr> "31", "31", "31", "31", "31", "29", "29", "29", "29~
## $ year               <chr> "2020", "2020", "2020", "2020", "2020", "2020", "20~
## $ day_of_week        <chr> "Saturday", "Saturday", "Saturday", "Saturday", "Sa~
```

Add new column to calculate each ride length in mins

```
all_trips_cleaned$ride_length <- difftime(all_trips_cleaned$ended_at,
                                          all_trips_cleaned$started_at)
```

Inspect the structure of the columns

```
str(all_trips_cleaned)
```

```
## tibble [4,317,599 x 15] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:4317599] "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396B54A15AC0DF" "44A4
##  $ rideable_type     : chr [1:4317599] "electric_bike" "electric_bike" "electric_bike" "electric_bik
##  $ started_at        : POSIXct[1:4317599], format: "2020-10-31 19:39:00" "2020-10-31 23:50:00" ...
##  $ ended_at          : POSIXct[1:4317599], format: "2020-10-31 19:57:00" "2020-11-01 00:04:00" ...
##  $ start_station_name: chr [1:4317599] "Lakeview Ave & Fullerton Pkwy" "Southport Ave & Waveland Ave
##  $ start_station_id  : chr [1:4317599] "313" "227" "102" "165" ...
##  $ end_station_name  : chr [1:4317599] "Rush St & Hubbard St" "Kedzie Ave & Milwaukee Ave" "Universi
##  $ end_station_id    : chr [1:4317599] "125" "260" "423" "256" ...
##  $ member_casual     : chr [1:4317599] "casual" "casual" "casual" "casual" ...
##  $ date              : Date[1:4317599], format: "2020-10-31" "2020-10-31" ...
##  $ month             : chr [1:4317599] "10" "10" "10" "10" ...
##  $ day               : chr [1:4317599] "31" "31" "31" "31" ...
##  $ year              : chr [1:4317599] "2020" "2020" "2020" "2020" ...
##  $ day_of_week       : chr [1:4317599] "Saturday" "Saturday" "Saturday" "Saturday" ...
##  $ ride_length       : 'difftime' num [1:4317599] 18 14 8 3 ...
##   ..- attr(*, "units")= chr "mins"
```

Convert "ride_length" from Factor to numeric so we can run calculations on the data

```
is.factor(all_trips_cleaned$ride_length)
```

```
## [1] FALSE
```

```
all_trips_cleaned$ride_length <- as.numeric(as.character(all_trips_cleaned$ride_length))
is.numeric(all_trips_cleaned$ride_length)
```

```
## [1] TRUE
```

Remove "bad" data and store in a new dataframe

```
all_trips_v2 <- all_trips_cleaned[!(all_trips_cleaned$start_station_name == "HQ QR" | all_trips_cleaned$
```

STEP 4: CONDUCT DESCRIPTIVE ANALYSIS

Descriptive analysis on ride_length (all figures in minutes)

```
mean(all_trips_v2$ride_length) #straight average (total ride length / rides)
```

```
## [1] 22.82869
```

```
median(all_trips_v2$ride_length) #midpoint number in the ascending array of ride lengths
```

```
## [1] 13
```

```
max(all_trips_v2$ride_length) #longest ride
```

```
## [1] 55944
```

```
min(all_trips_v2$ride_length) #shortest ride
```

```
## [1] 1
```

```
# You can condense the four lines above to one line using summary() on the specific attribute
summary(all_trips_v2$ride_length)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     1.00     7.00    13.00    22.83    23.00 55944.00
```

Compare members and casual users

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = mean)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                 33.73528
## 2                     member                 13.82519
```

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = median)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                       17
## 2                     member                       10
```

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = max)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                    55944
## 2                     member                     9558
```

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = min)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual                       1
## 2                     member                       1
```

See the average ride time by each day for members vs casual users

```
# Arranging the days of the week accordingly
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "

aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, FUN = mean)
```

```
##    all_trips_v2$member_casual all_trips_v2$day_of_week all_trips_v2$ride_length
## 1                      casual                   Sunday                39.32918
## 2                      member                   Sunday                15.79215
## 3                      casual                   Monday                33.10689
## 4                      member                   Monday                13.24227
## 5                      casual                  Tuesday                30.37187
## 6                      member                  Tuesday                13.02253
## 7                      casual                Wednesday                29.18995
## 8                      member                Wednesday                13.07650
## 9                      casual                 Thursday                28.98451
## 10                     member                 Thursday                12.96861
## 11                     casual                   Friday                32.41452
## 12                     member                   Friday                13.60256
## 13                     casual                 Saturday                36.23859
## 14                     member                 Saturday                15.47130
```

Analyze ridership data by type and weekday

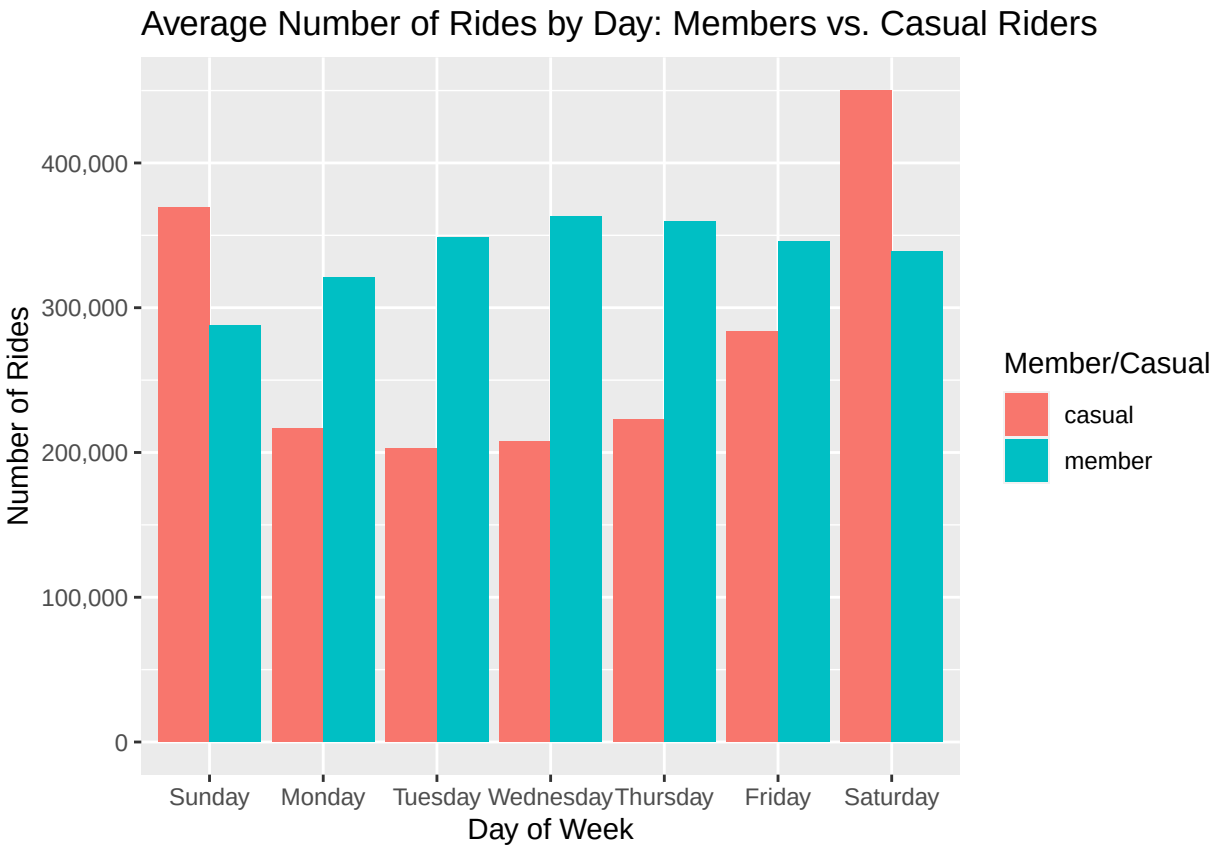
```
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%  #creates weekday field using wday()
  group_by(member_casual, weekday) %>%  #groups by usertype and weekday
  summarise(number_of_rides = n()                              #calculates the number of rides and averag
            ,average_duration = mean(ride_length)) %>%         # calculates the average duration
  arrange(member_casual, weekday)                                # sorts
```

```
## `summarise()` has grouped output by 'member_casual'. You can override using the `.groups` argument.
```

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual weekday number_of_rides average_duration
##    <chr>         <ord>             <int>            <dbl>
##  1 casual        Sun              379738             38.9
##  2 casual        Mon              217785             33.4
##  3 casual        Tue              202323             30.2
##  4 casual        Wed              207359             29.4
##  5 casual        Thu              221228             28.9
##  6 casual        Fri              277051             32.4
##  7 casual        Sat              446968             36.3
##  8 member        Sun              292528             15.8
##  9 member        Mon              321081             13.3
## 10 member        Tue              348428             13.0
## 11 member        Wed              362977             13.1
## 12 member        Thu              358888             13.0
## 13 member        Fri              342739             13.6
## 14 member        Sat              338506             15.4
```

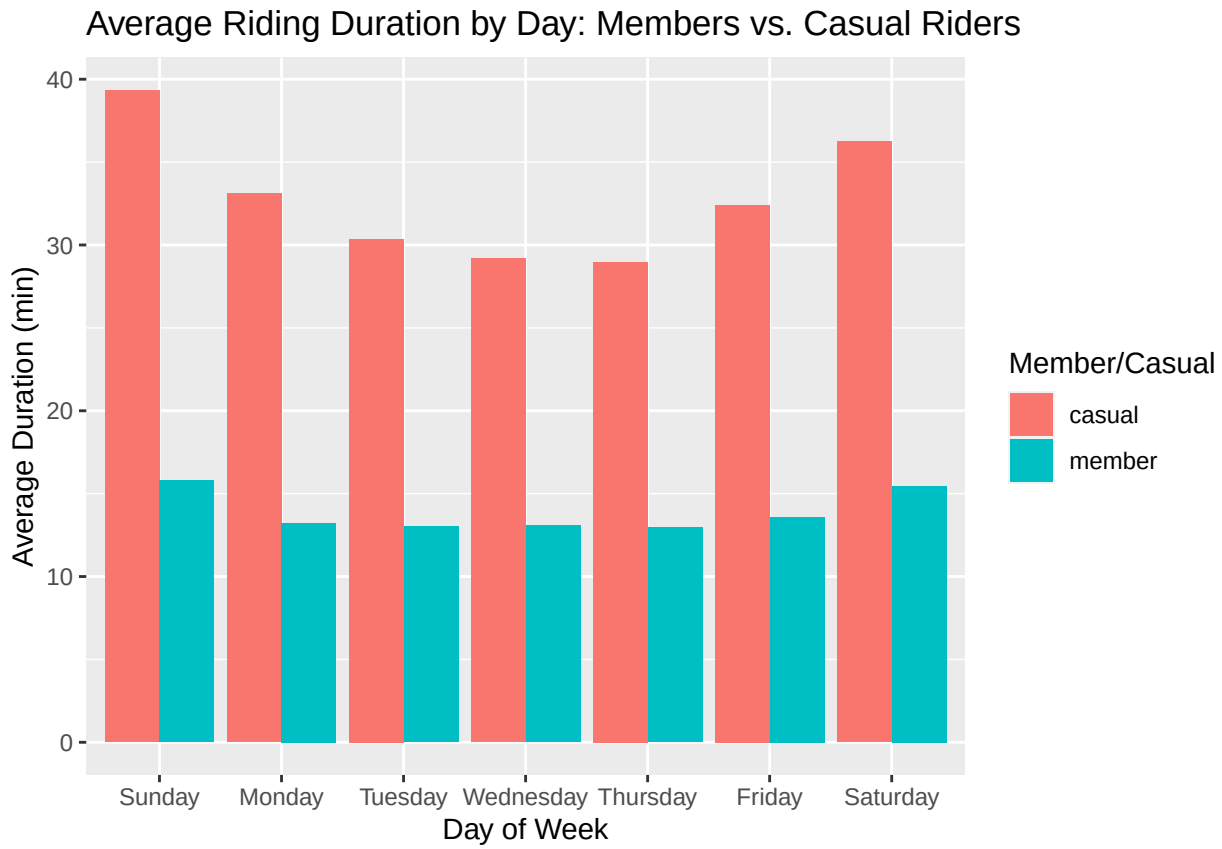
Visual for number of rides grouped by rider type

```
all_trips_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(number_of_rides = n(), .groups = 'drop') %>%
  #arrange(member_casual, day_of_week) %>%
  ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") + scale_y_continuous(labels = scales::comma) +
  labs(x = "Day of Week", y = "Number of Rides", fill = "Member/Casual",
       title = "Average Number of Rides by Day: Members vs. Casual Riders")
```



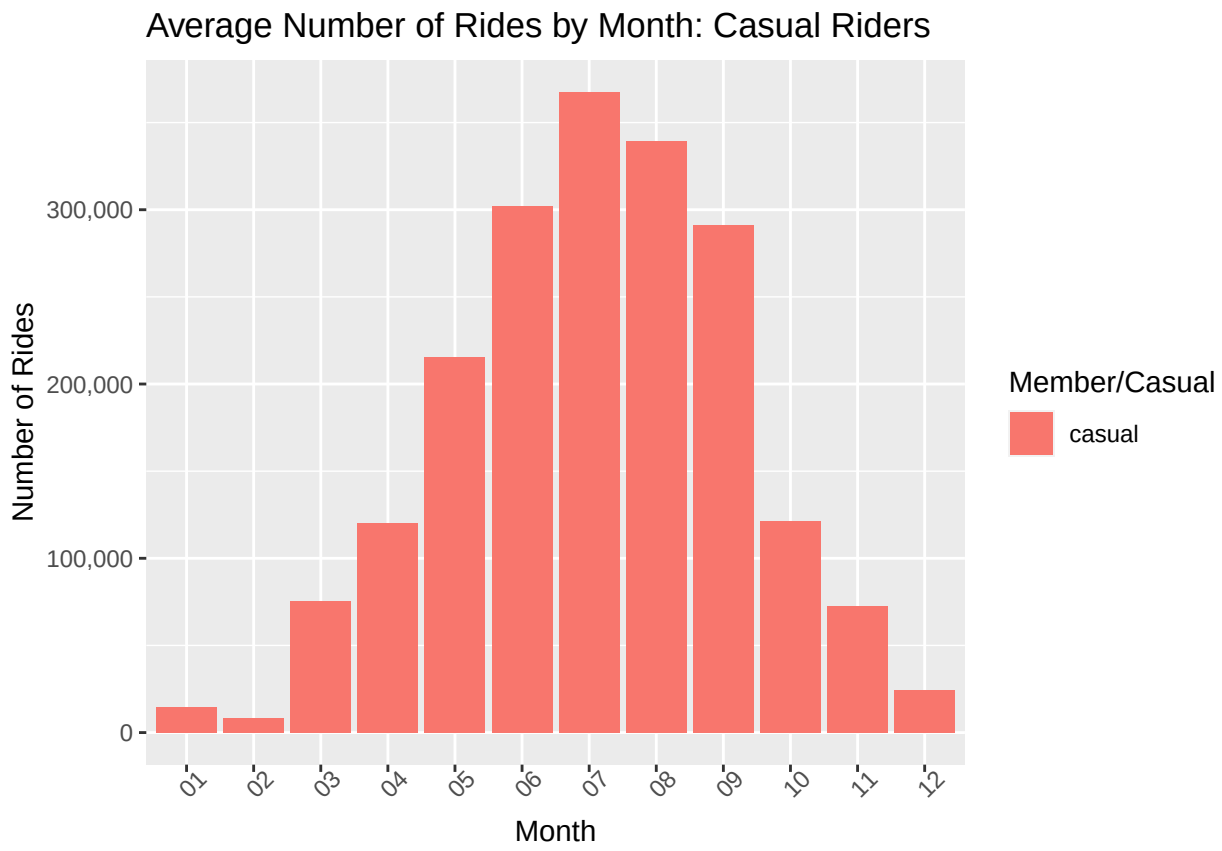
Visual for average duration

```
all_trips_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarise(average_duration = mean(ride_length), .groups = 'drop') %>%
  #arrange(member_casual, day_of_week) %>%
  ggplot(aes(x = day_of_week, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(x = "Day of Week", y = "Average Duration (min)",
       fill = "Member/Casual",
       title = "Average Riding Duration by Day: Members vs. Casual Riders")
```



Average Number of Rides by Month

```
all_trips_v2 %>%
  group_by(month, member_casual) %>%
  summarize(number_of_rides = n(), .groups = 'drop') %>%
  filter(member_casual == 'casual') %>%
  drop_na() %>%
  ggplot(aes(x = month, y = number_of_rides, fill = member_casual)) +
  geom_bar(position = 'dodge', stat = 'identity') + scale_y_continuous(labels = scales::comma) +
  theme(axis.text.x = element_text(angle = 45)) +
  labs(x = "Month", y = "Number of Rides",
       fill = "Member/Casual",
       title = "Average Number of Rides by Month: Casual Riders")
```



4.1 MORE DESCRIPTIVE DATA ANALYSIS

```
# Combine start and end stations
# Removing entries with no station name
# Separate the data frame by rider type
all_stations <- bind_rows(data.frame("stations" = all_trips_v2$start_station_name,
                                     "member_casual" = all_trips_v2$member_casual),
                          data.frame("stations" = all_trips_v2$end_station_name,
                                     "member_casual" = all_trips_v2$member_casual))
all_stations_v2 <- all_stations[!(all_stations$stations == "" | is.na(all_stations$stations)),]
all_stations_member <- all_stations_v2[all_stations_v2$member_casual == 'member',]
all_stations_casual <- all_stations_v2[all_stations_v2$member_casual == 'casual',]

# Get the top 10 popular stations all, members, and casual riders
top_10_station <- all_stations_v2 %>%
  group_by(stations) %>%
  summarise(station_count = n()) %>%
  arrange(desc(station_count)) %>%
  slice(1:10)

top_10_station_member <- all_stations_member %>%
  group_by(stations) %>%
  summarise(station_count = n()) %>%
  arrange(desc(station_count)) %>%
  head(n=10)
```

```r
top_10_station_casual <- all_stations_casual %>%
  group_by(stations) %>%
  summarise(station_count = n()) %>%
  arrange(desc(station_count)) %>%
  head(n=10)

# Comparing general bike type preference between members and casual riders
all_trips_v2 %>%
  group_by(rideable_type, member_casual) %>%
  summarize(number_of_rides = n(), .groups = 'drop')
```

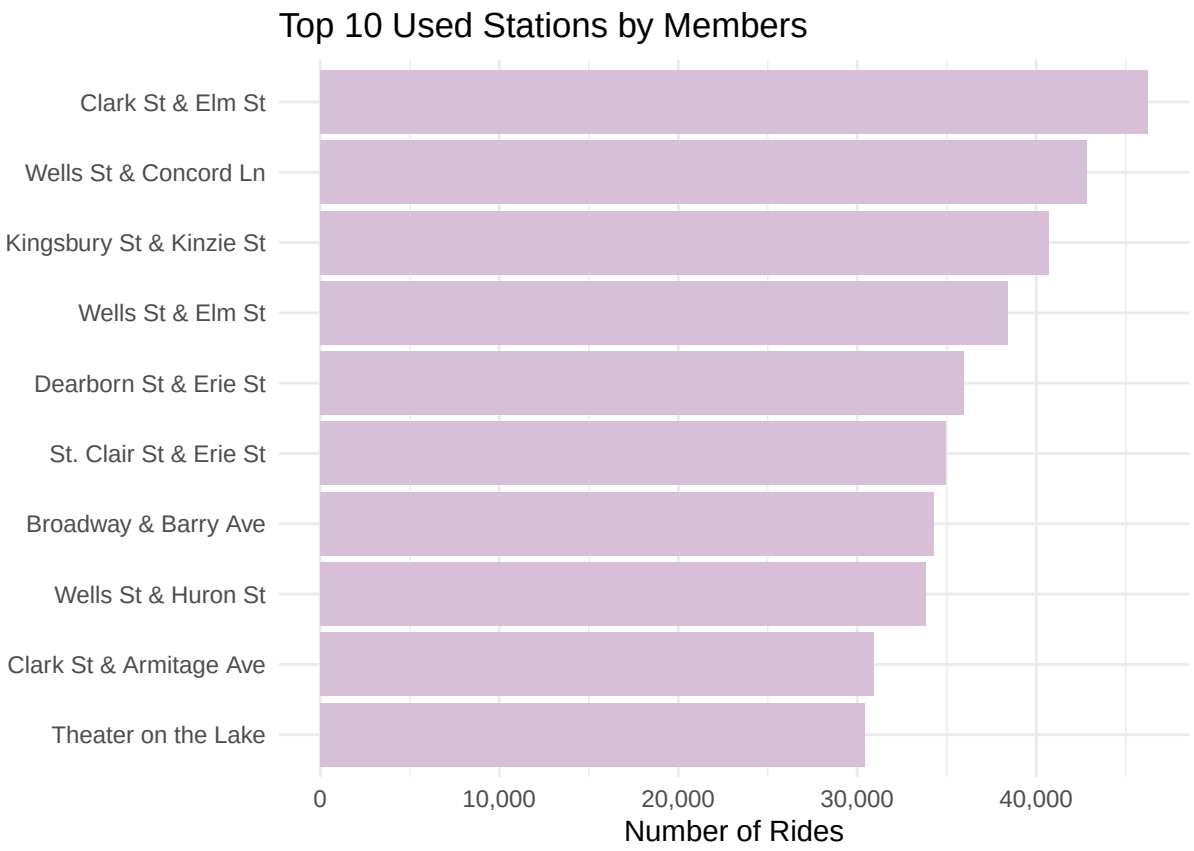
```
## # A tibble: 6 x 3
##   rideable_type member_casual number_of_rides
##   <chr>         <chr>                  <int>
## 1 classic_bike  casual               1109838
## 2 classic_bike  member               1612212
## 3 docked_bike   casual                405030
## 4 docked_bike   member                264632
## 5 electric_bike casual                437584
## 6 electric_bike member                488303
```

```r
# average number of rides by hour (casual riders)
all_trips_v2$started_at_hour <- as.POSIXct(all_trips_v2$started_at, "%Y-%m-%d %H:%M")
str(all_trips_v2)
```

```
## tibble [4,317,599 x 16] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:4317599] "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396B54A15AC0DF" "44A4
##  $ rideable_type     : chr [1:4317599] "electric_bike" "electric_bike" "electric_bike" "electric_bik
##  $ started_at        : POSIXct[1:4317599], format: "2020-10-31 19:39:00" "2020-10-31 23:50:00" ...
##  $ ended_at          : POSIXct[1:4317599], format: "2020-10-31 19:57:00" "2020-11-01 00:04:00" ...
##  $ start_station_name: chr [1:4317599] "Lakeview Ave & Fullerton Pkwy" "Southport Ave & Waveland Ave
##  $ start_station_id  : chr [1:4317599] "313" "227" "102" "165" ...
##  $ end_station_name  : chr [1:4317599] "Rush St & Hubbard St" "Kedzie Ave & Milwaukee Ave" "Universi
##  $ end_station_id    : chr [1:4317599] "125" "260" "423" "256" ...
##  $ member_casual     : chr [1:4317599] "casual" "casual" "casual" "casual" ...
##  $ date              : Date[1:4317599], format: "2020-10-31" "2020-10-31" ...
##  $ month             : chr [1:4317599] "10" "10" "10" "10" ...
##  $ day               : chr [1:4317599] "31" "31" "31" "31" ...
##  $ year              : chr [1:4317599] "2020" "2020" "2020" "2020" ...
##  $ day_of_week       : Ord.factor w/ 7 levels "Sunday"<"Monday"<..: 7 7 7 7 7 5 5 5 5 4 ...
##  $ ride_length       : num [1:4317599] 18 14 8 3 16 7 14 15 12 3 ...
##  $ started_at_hour   : POSIXct[1:4317599], format: "2020-10-31 19:39:00" "2020-10-31 23:50:00" ...
```

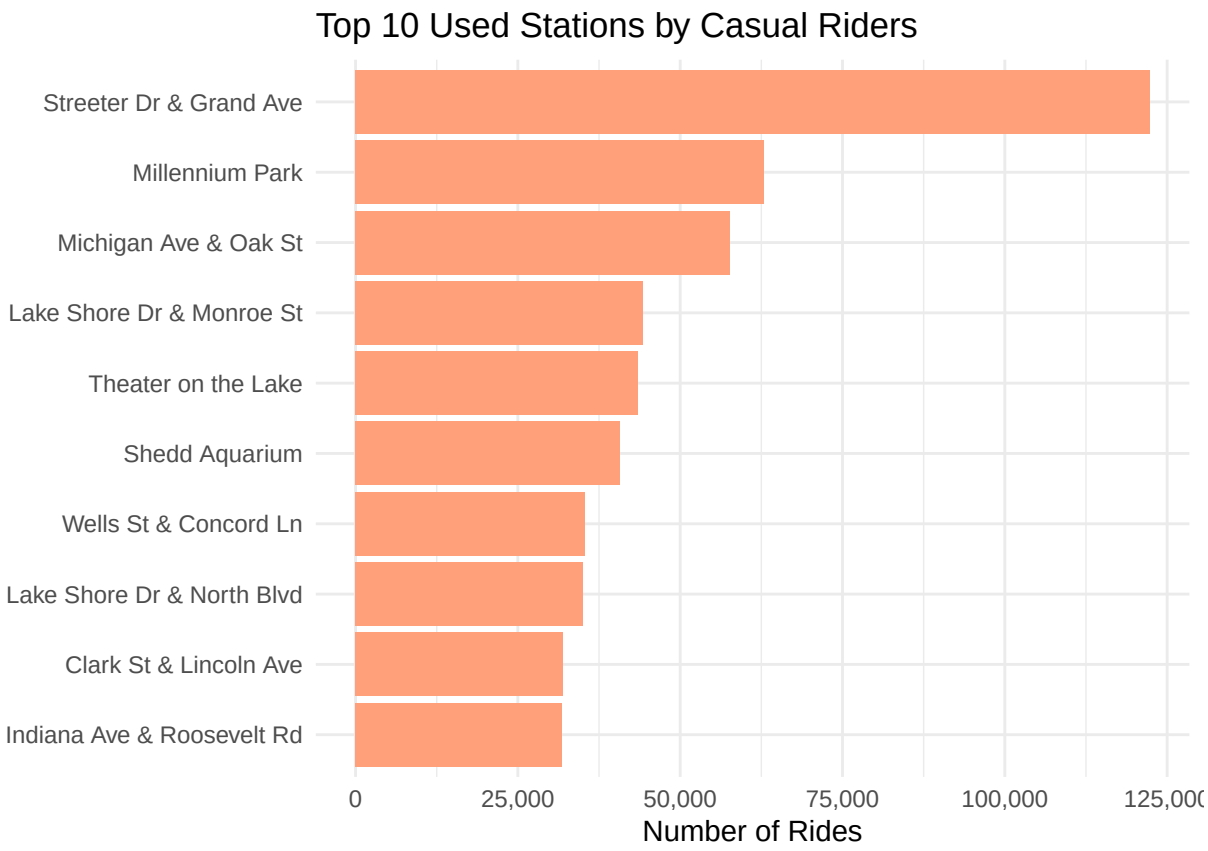
Visual for Top 10 Used Stations by Members

```r
ggplot(data = top_10_station_member) +
  geom_col(aes(x = reorder(stations, station_count), y = station_count), fill = "thistle") +
  labs(title = "Top 10 Used Stations by Members", y = "Number of Rides", x = "") +
  scale_y_continuous(labels = scales::comma) +
  coord_flip() +
  theme_minimal()
```



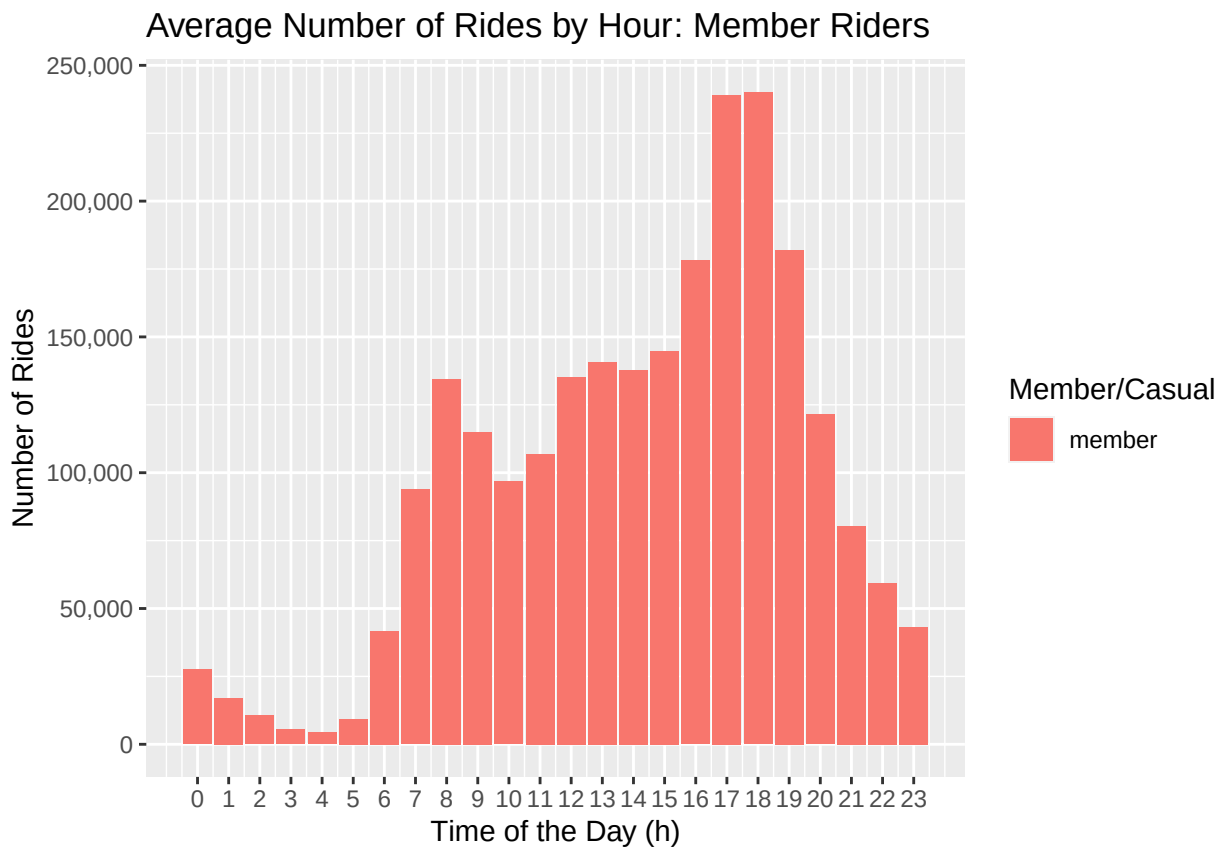
Visual for Top 10 Used Stations by Casual Members

```
ggplot(data = top_10_station_casual) +
  geom_col(aes(x = reorder(stations, station_count), y = station_count), fill = "lightsalmon") +
  labs(title = "Top 10 Used Stations by Casual Riders", x = "", y = "Number of Rides") +
  scale_y_continuous(labels = scales::comma) +
  coord_flip() +
  theme_minimal()
```



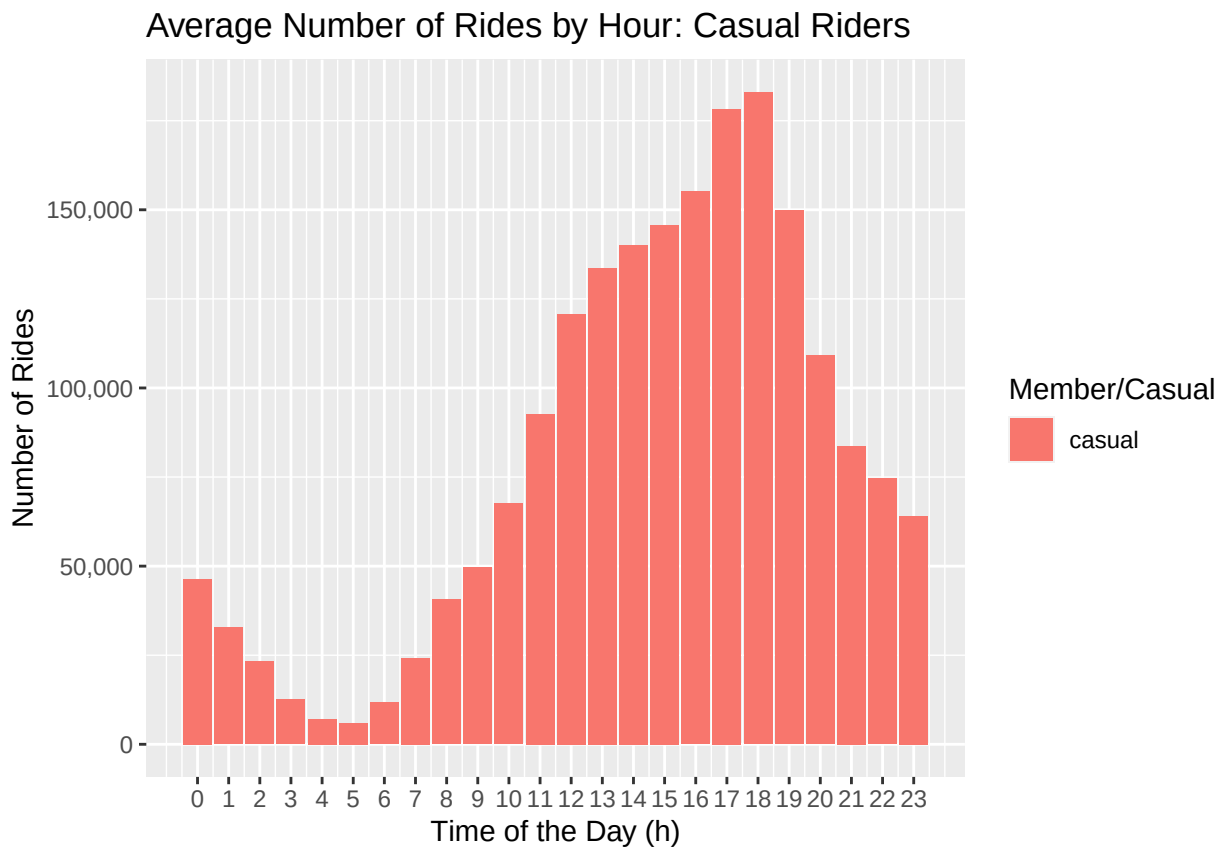
Visual for Average Number of Rides by Hour (Member riders)

```
all_trips_v2 %>%
  filter(member_casual == 'member') %>%
  group_by(hour_of_day = hour(round_date(started_at_hour, 'hour'))) %>%
  group_by(hour_of_day, member_casual) %>%
  summarize(number_of_rides = n(), .groups = 'drop') %>%
  arrange(-number_of_rides) %>%
  ggplot(aes(x = hour_of_day, y = number_of_rides, fill = member_casual)) +
  geom_bar(position = 'dodge', stat = 'identity') + scale_y_continuous(labels = scales::comma) +
  scale_x_continuous(breaks = c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23)) +
  labs(x = "Time of the Day (h)", y = "Number of Rides",
       fill = "Member/Casual",
       title = "Average Number of Rides by Hour: Member Riders")
```



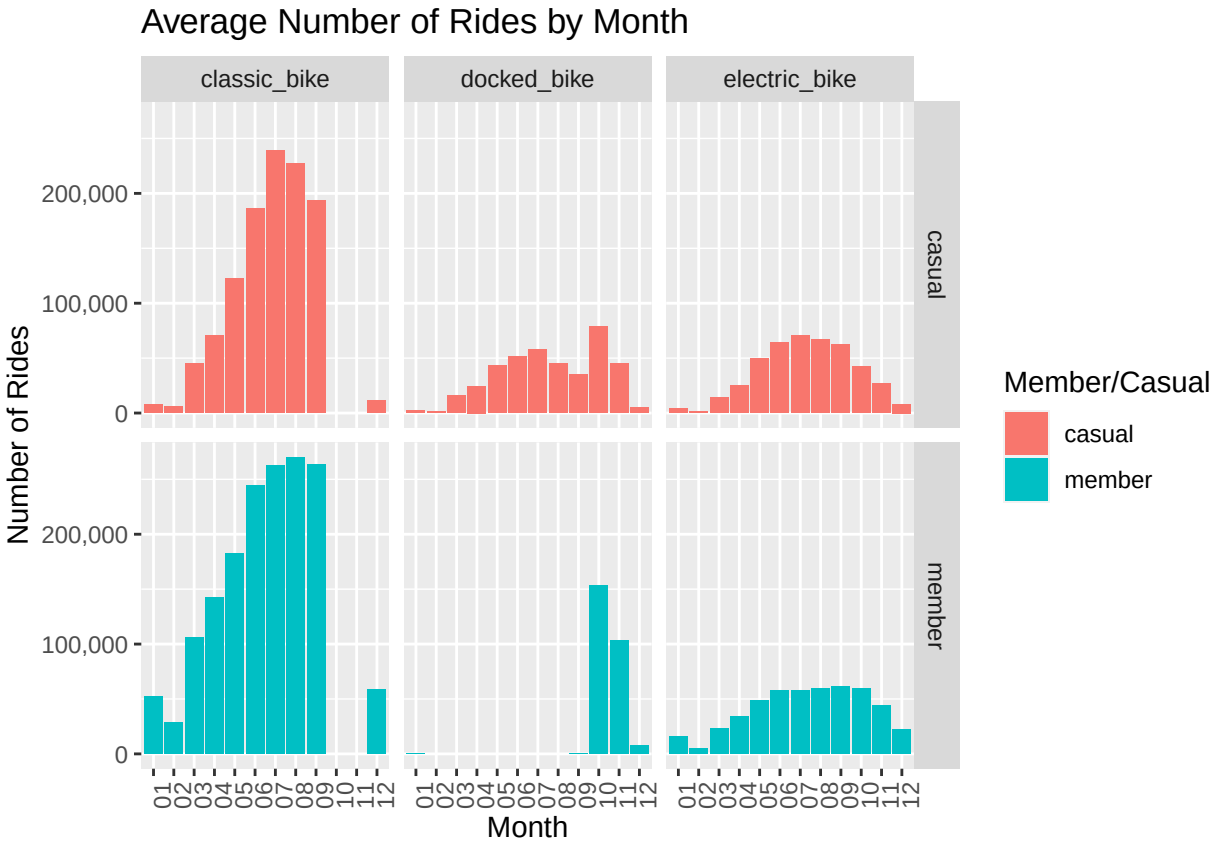
Visual for average number of rides by hour (casual riders)

```
options(repr.plot.width = 12, repr.plot.height = 8)
all_trips_v2 %>%
  filter(member_casual == 'casual') %>%
  group_by(hour_of_day = hour(round_date(started_at_hour, 'hour'))) %>%
  group_by(hour_of_day, member_casual) %>%
  summarize(number_of_rides = n(), .groups = 'drop') %>%
  arrange(-number_of_rides) %>%
  ggplot(aes(x = hour_of_day, y = number_of_rides, fill = member_casual)) +
  geom_bar(position = 'dodge', stat = 'identity') + scale_y_continuous(labels = scales::comma) +
  scale_x_continuous(breaks = c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23)) +
  labs(x = "Time of the Day (h)", y = "Number of Rides",
       fill = "Member/Casual",
       title = "Average Number of Rides by Hour: Casual Riders")
```



Visual for usage of different bikes by rider type (separated)

```
options(repr.plot.width = 14, repr.plot.height = 10)
all_trips_v2 %>%
  group_by(month, member_casual, rideable_type) %>%
  summarize(number_of_rides = n(), .groups = 'drop') %>%
  drop_na() %>%
  ggplot(aes(x = month, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  scale_y_continuous(labels = scales::comma) +
  facet_grid(member_casual~rideable_type) +
  labs(x = "Month", y = "Number of Rides", fill = "Member/Casual",
       title = "Average Number of Rides by Month") +
  theme(axis.text.x = element_text(angle = 90))
```



Key takeaways:

- The average ride duration is higher for casual riders for any day of the week.
- Both members and casual riders preferred docked bikes, while the classic bike is the least popular bike type.
- Streeter Dr & Grand Ave, Lake Shore Dr & Monroe St, and Millennium Park are casual riders' top three start stations.
- Casual riders ride more during the weekends.

Recommendations

- Giving incentives or rewards for achieving members' milestones to attract casual riders to become members.
- Offer occasional membership discount to new riders on summer and holiday weekends
- Partner with local businesses within the top used stations for casual riders targeting 1) local casual riders, 2) frequent visitors (commuters) to the businesses.